Math 343 / 643 Fall 2024
Midterm Examination Two Solutions

Professor Adam Kapelner

April 9, 2024

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

____ ________________

signature date

Instructions

This exam is 75 minutes and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 Consider the following matrix of constant measurements:

$$\boldsymbol{X} := [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ \ldots \ \mid \boldsymbol{x}_{\cdot p}]$$

with column indices $0, 1, \ldots, p$ and row indices $1, 2, \ldots, n$. We assume also a continuous (real-valued) response model which is linear in these measurements, i.e.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}.$$

For the error term, we assume the "core assumption",

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0}_n,\, \sigma^2 \boldsymbol{I}_n\right).$$

And for our estimator of $\boldsymbol{\beta}$, we choose:

$$\boldsymbol{B} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

(a) [5 pt / 5 pts] In the "linear response model assumption line" above, list the scalar parameters (if the parameter is a vector, list the scalar entries). If there are no parameters, write "none".

$\beta_0, \beta_1, \ldots, \beta_p$

(b) [5 pt / 10 pts] In the "core assumption" line above, list the scalar parameters (if the parameter is a vector, list the scalar entries). If there are no parameters, write "none".

σ^2

(c) [5 pt / 15 pts] In the "our estimator" line above, list the scalar parameters (if the parameter is a vector, list the scalar entries). If there are no parameters, write "none".

none

(d) [10 pt / 25 pts] Show that $\boldsymbol{B}$ is unbiased.

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{B}\right] &= \mathbb{E}\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}})\right] \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathbb{E}\left[\boldsymbol{\mathcal{E}}\right] \\
&= \boldsymbol{\beta} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathbf{0}_n \\
&= \boldsymbol{\beta}
\end{aligned}$$

Now choose the following estimator instead

$$\boldsymbol{B}_{\text{ridge}} := \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_{p+1}\right)^{-1} \boldsymbol{X}^T\boldsymbol{Y} \quad \text{where } \lambda > 0.$$

(e) [6 pt / 31 pts] Which of these two will be larger: $||\boldsymbol{B}||^2$ or $||\boldsymbol{B}_{\text{ridge}}||^2$?

$||\boldsymbol{B}||^2$

Problem 2 This problem will analyze data from a study that investigated tooth cell growth (in length) in guinea pigs by vitamin C dose (0.5, 1 or 2mg/d) and delivery method (OJ = orange juice or VC = vitamin capsule). Here is the results of an OLS model fit to both dose and delivery method:

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.2725     1.2824   7.231 1.31e-09 ***
deliveryVC        -3.7000     1.0936  -3.383   0.0013 **
dose               9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,        Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Unless otherwise noted, assume that $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\boldsymbol{0}_n, \sigma^2\boldsymbol{I}_n\right)$. Let $\boldsymbol{X}$ denote the design matrix for this linear regression. The following quantities may be useful:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 60 & 30 & 70 \\ 30 & 30 & 30 \\ 70 & 35 & 105 \end{bmatrix}, \quad \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \begin{bmatrix} 0.09 & -0.03 & -0.05 \\ -0.03 & 0.07 & 0.00 \\ -0.05 & 0.00 & 0.04 \end{bmatrix},$$

(a) [5 pt / 36 pts] What is the sample size n of this dataset? $57 + (2 + 1) = 60$

(b) [5 pt / 41 pts] Test $H_0 : \beta_{\text{dose}} = 0$ at $\alpha = 5\%$.

This was done for us in the results printed above. The p-value is $6.31 \times 10^{-16} < \alpha = 0.05$ hence H_0 is rejected.

(c) [5 pt / 46 pts] Create a 95% CI for the effect of delivery being a vitamin capsule instead of orange juice. The t-value to use in this computation is 2.00.

$[-3.70 \pm 2.00 \cdot 1.09] = [-5.88, -1.52]$

(d) [5 pt / 51 pts] Run the omnibus test at $\alpha = 5\%$.

This was done for us in the results printed above. The p-value is $8.716 \times 10^{-16} < \alpha = 0.05$ hence H_0 is rejected.

(e) [10 pt / 61 pts] For a guinea pig who was given orange juice and 1mg of vitamin C, provide a 95% CI for the guinea pig's tooth cell growth. The t-value to use in this computation is 2.00.

$$
\begin{aligned}
CI_{y_\star, 95\%} &= \left[\hat{y} \pm t_{1-\alpha/2, n-(p+1)} \cdot s_e \sqrt{1 + \boldsymbol{x}_\star^\top \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_\star}\right] \\
&= \left[(9.2725 + 9.7636) \pm 2.00 \cdot 4.236 \sqrt{1 + [1\ 0\ 1] \begin{bmatrix} 0.09 & -0.03 & -0.05 \\ -0.03 & 0.07 & 0.00 \\ -0.05 & 0.00 & 0.04 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}}\right] \\
&= \left[19.0361 \pm 2 \cdot 4.236 \sqrt{1.03}\right] \\
&= [10.438, 27.634]
\end{aligned}
$$

(f) [10 pt / 71 pts] Assuming independence of errors and homoskedasticity of errors, test $H_0 : \beta_{\text{dose}} = 10$ at $\alpha = 5\%$.

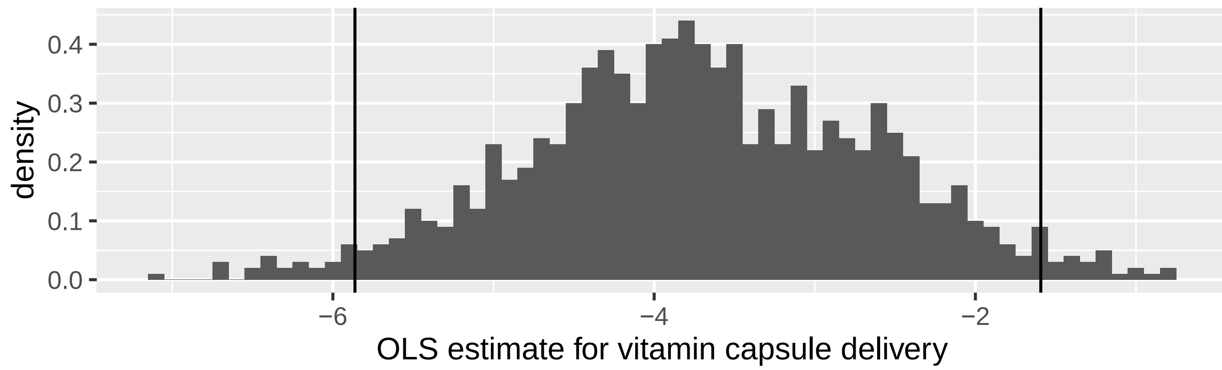
We can use the Wald test (but we cannot use the t-test):

$$
\frac{B_j - \beta_j}{\text{SE}[B_j]} \mathrel{\dot\sim} \mathcal{N}(0, 1) \quad \Rightarrow \quad \frac{9.7636 - 10}{0.8768} = -0.2696 \in [-1.96, 1.96]
$$

Thus we fail to reject H_0.

(g) [5 pt / 76 pts] Is there enough information here to test $H_0 : \beta_{\text{dose}} = 0$ at $\alpha = 5\%$ if we assume independent, mean-centered and heteroskedastic errors? Yes / $\boxed{\text{No}}$

In reality, we are unsure of the distribution of the errors but we know that due to the way the data was sampled, we are guaranteed that the errors are independent. Hence use a bootstrap. We are interested in inference for the effect of delibery being a vitamin capsule instead of orange juice. The top of the following page shows the result of 1,000 boostrap samples where each time, the OLS for this effect was computed. The vertical lines indicate the empirical 2.5%ile and 97.5%ile.



(h) [5 pt / 81 pts] Using the boostrap samples, test $H_0 : \beta$ for vitamin capsule delivery is zero at $\alpha = 5\%$.

H_0 is rejected because $0 \notin CI_{\beta, 1-\alpha} = [-5.9, -1.7]$.

Problem 3 This problem will analyze data from a study that investigated the number of yarn breaks by two features: type of yarn wool (A or B) and amount of tension (L = low, M = medium, H = high). Since the response being modeled is a count, we choose a negative binomial model which is more flexible than a Poission model. We parameterize the negative binomial with parameters r, θ where its expectation is θ. We link θ to the two features using the link $\theta_i = e^{\boldsymbol{x}_i \boldsymbol{\beta}}$ for $i = 1, \ldots, n$. Below is the summary for the inference of $\boldsymbol{\beta}$ for both features (the inference for r is omitted). We also display the log-likelihood of this model.

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.6734     0.0979  37.520  < 2e-16 ***
woolB        -0.1862     0.1010  -1.844   0.0651 .
tensionM     -0.2992     0.1217  -2.458   0.0140 *
tensionH     -0.5114     0.1237  -4.133 3.58e-05 ***

'log Lik.' -199.3819 (df=5)
```

Here are outputs for two other models:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.43518    0.08071  42.562   <2e-16 ***
woolB       -0.20599    0.11533  -1.786   0.0741 .

'log Lik.' -206.9874 (df=3)
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.59426    0.08715  41.243  < 2e-16 ***
tensionM    -0.32132    0.12557  -2.559   0.0105 *
tensionH    -0.51849    0.12739  -4.070  4.7e-05 ***

'log Lik.' -201.0109 (df=4)
```

Here are some values of the inverse CDF of the χ^2_{df} distribution:

	Probability less than the critical value				
df	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322

(a) [4 pt / 85 pts] Is the parameter r a nuisance parameter? $\boxed{\text{Yes}}$ / No

(b) [6 pt / 91 pts] Calculate the likelihood ratio test statistic for the test that tension has no effect on number of yarn breaks.

$$\hat{\hat{\Lambda}} = 2(\ln(\mathcal{L}_{\text{full}}) - \ln(\mathcal{L}_{\text{reduced}})) = 2(-199.3819 - -206.9874) = 15.211$$

(c) [3 pt / 94 pts] Test the null hypothesis that tension has no effect on number of yarn breaks at $\alpha = 5\%$. Justify your answer.

The difference in df of full to reduced is 2. Hence the likelihood ratio estimator is asymptotically distributed as χ^2_2. At $\alpha = 5\%$, according to the table, the critical cutoff value is 5.991. We found the LRT statistic to be 15.211 which is greater. Hence, we reject H_0.

(d) [6 pt / 100 pts] The maximum likelihood estimate of r is 9.94. Given a piece of yarn with wool type B and low tension, predict the number of yarn breaks it will have to the nearest number of yarn breaks.

$$y_\star = \text{round}(e^{\boldsymbol{x}_\star \boldsymbol{b}}) = \text{round}(e^{3.6734 + -0.1862}) = \text{round}(32.6943) = 33$$